Sampling representative years for a TSO in a climate simulation of 200 years.

Jean Thorey[1]

[1] RTE, Paris, FRANCE. jean.thorey @ rte-france.com

Introduction

We have several datasets of **200 years of hourly gridded data** under future climate. Because power-flow simulations for 200 years are computationally costly, our objective is to find a small number of representative years.

Statistical models based on historical data **convert weather variables to energy-related variables** : consumption, and wind and photovoltaic (PV) capacity factors. Assumptions on the evolution of electricity generation and consumption (population growth, new wind power plants, technology evolution, electric mobility…) come from our **Long-Term Adequacy Report** [1].

The sample of roughly 10 years should be close to the full dataset. We resort to the definition of constraints to filter sample candidates.

Sample scoring

For each variable and location, we compute several scores between the sample u and the full distribution v :

- e.g. $|\widehat{\mu}^{austria}_{t2m} - \mu^{austria}_{t2m}|$.
- A year is a distribution of timesteps.
- A sample u of years is a mixture of distributions $f_u = \sum_{j=1}^{N} u_j f_j$.
- A sample u is compared to the uniform mixture v.

Table 1: Score definitions.

Score	Formula	Speed
Mean value	$\lvert\widehat{\mu} - \mu\rvert = \lvert u^\top m - v^\top m\rvert$	Fast
Energy score	$u^\top A v - \frac{1}{2} u^\top A u - \frac{1}{2} v^\top A v$	Fast
Quantile MAE	$\frac{1}{N_q} \sum_\alpha \lvert q(\alpha) - \hat{q}(\alpha)\rvert$	Slow
Quantile MAE peak	$\frac{1}{N_p} \sum_{\alpha > \alpha_{peak}} \lvert q(\alpha) - \hat{q}(\alpha)\rvert$	Slow
Quantile error 99^{th}	$\lvert q(\alpha^{99}) - \hat{q}(\alpha^{99})\rvert$	Slow

For the energy score, we precompute the matrix A with :
- $ES = E(||X - Y||) - \frac{1}{2} E(||X - X'||) - \frac{1}{2} E(||Y - Y'||)$.

Sampling strategies

We want $\forall i$, $score_i < M_i$. Lowering M_i increases difficulty.

To generate a new sample, we can use :
- random sampling,
- random sampling based on a clustering of years [2],
- local search optimization.

To save computation time :
- we compute costly scores only if cheap scores are satisfying.

In practice, we tested 10^7 samples for each experiment.

Experiment 1 : weather

Data set 1 = weather-related :
- Climate projections 2050-2080 RCP 8.5.
- Weather variables : temperature, wind speed and solar radiation.
- 7 models $\rightarrow$ 210 years.
- 5 score types $\times$ 25 regional aggregates $\times$ 3 variables = 375 scores.

What about multi-model variability ?
- We want to avoid overfitting, i.e. M_i too low.
- We want $M_i >$ multi-model variability.
- With two samples P_{90}, P'_{90} of 90 years (2×3 models among 7),
- we set $M_i = \gamma \times \max(score_i(P_{90}, P'_{90}))$ with $\gamma = 1.5$.
- Numerical experiments for $\binom{90}{10} \approx 10^{12}$.

We illustrate :
- the large intermodel variability (large M_i).
- the importance of the sample size (number of years).
- the filtering efficiency of the number of constraints.
- the approximation error of a selected sample.

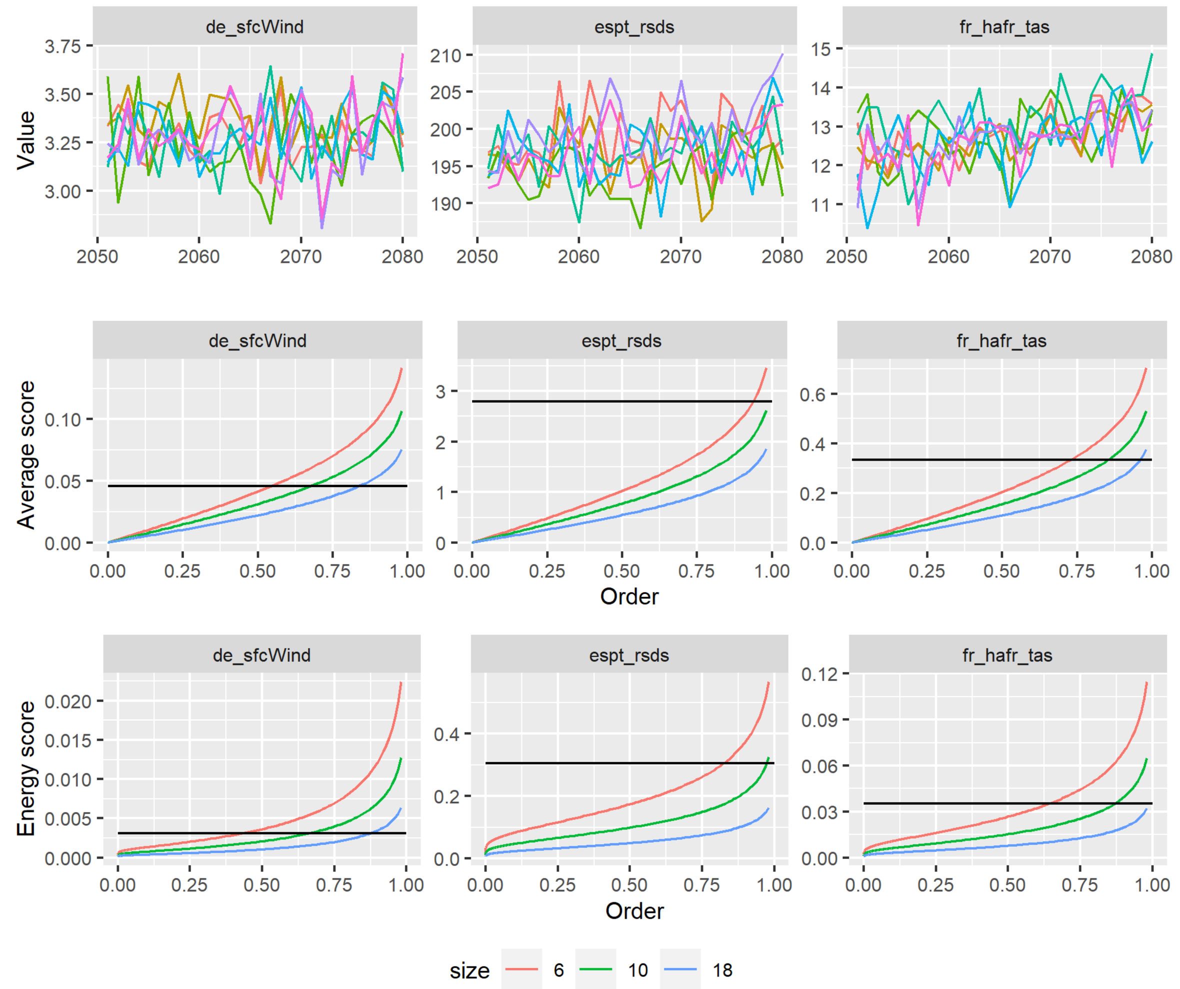

Figure 1: Top line : series of annual average values. Middle and bottom line : sorted scores of random samples according to the sample size (number of years). The black line indicates the threshold M_i.

Table 2: Number of random samples among 10^7 verifying $\forall i$ $score_i < M_i$. The number 25000 is an estimation (last row).

Size	avg	energy	avg+energy	avg+energy+quantile
6	2482	241	28	0
10	54347	128814	17631	9
18	727397	2790931	634804	25000

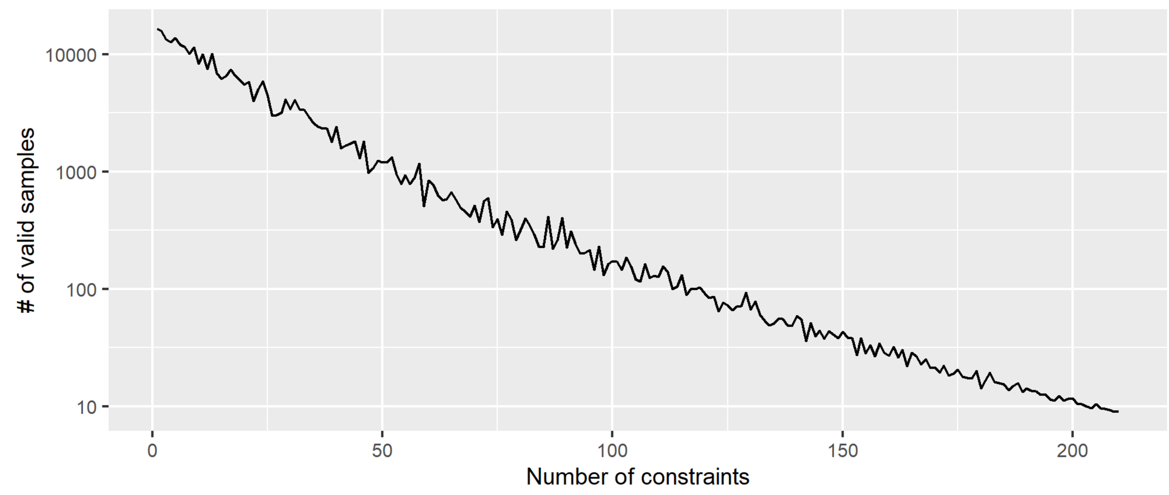

Figure 2: Number of satisfying random samples depending on the number of randomly picked constraints of quantile score, for samples of 10 years.

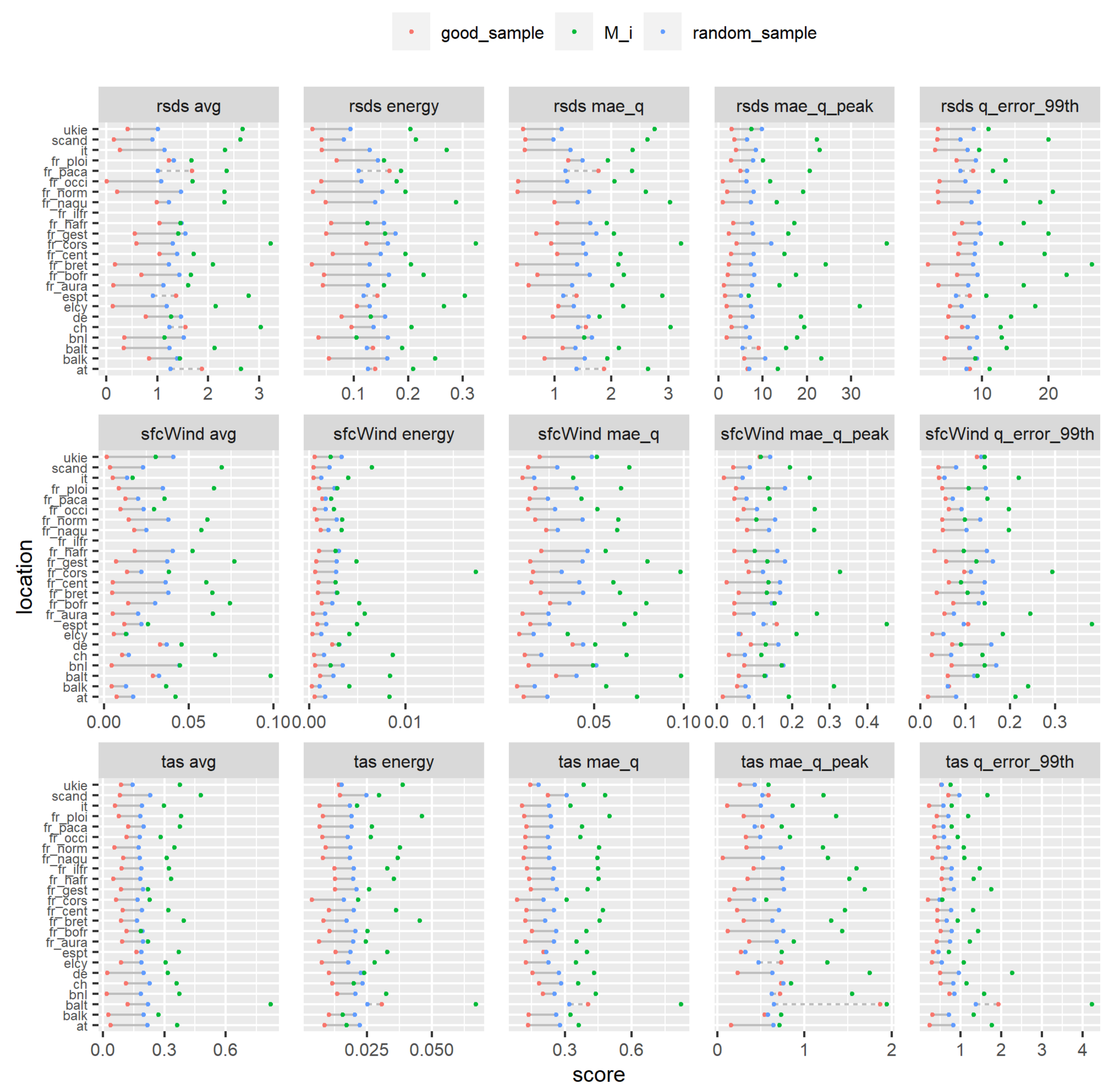

Figure 3: Scores of a satisfying sample vs. average scores of random samples.

References

[1] RTE. Long-Term Adequacy Report. URL: https://rte-futursenergetiques2050.com/

[2] Gabor J Szekely and Maria L Rizzo. “Hierarchical clustering via joint between-within distances: ExtendingWard’s minimum variance method.” In: Journal of classification 22.2 (2005), pp. 151–183.

[3] RTE. Antares simulator. URL: https://antares-simulator.org.

Experiment 2 : energy

Data set 2 = energy-related :
- 200 years of Arpege-Climate model for 2050 under RCP 8.5.
- Weather $\rightarrow$ loads and renewables $\rightarrow$ generations $\rightarrow$ power flows.
- Antares simulator [3] optimizes the production unit commitment.
- 2 socio-economic scenarios.
- Energy variables : regional/national/continental net load, flows . . .
- Operational use of the sample of 10 years.

No other data ? optimize !
- Custom score thresholds M_i.
- Numerical experiments for $\binom{200}{10} \approx 10^{16}$.

We illustrate :
- the difference between the sampled and the full distribution.
- the approximation error of a satisfying sample.

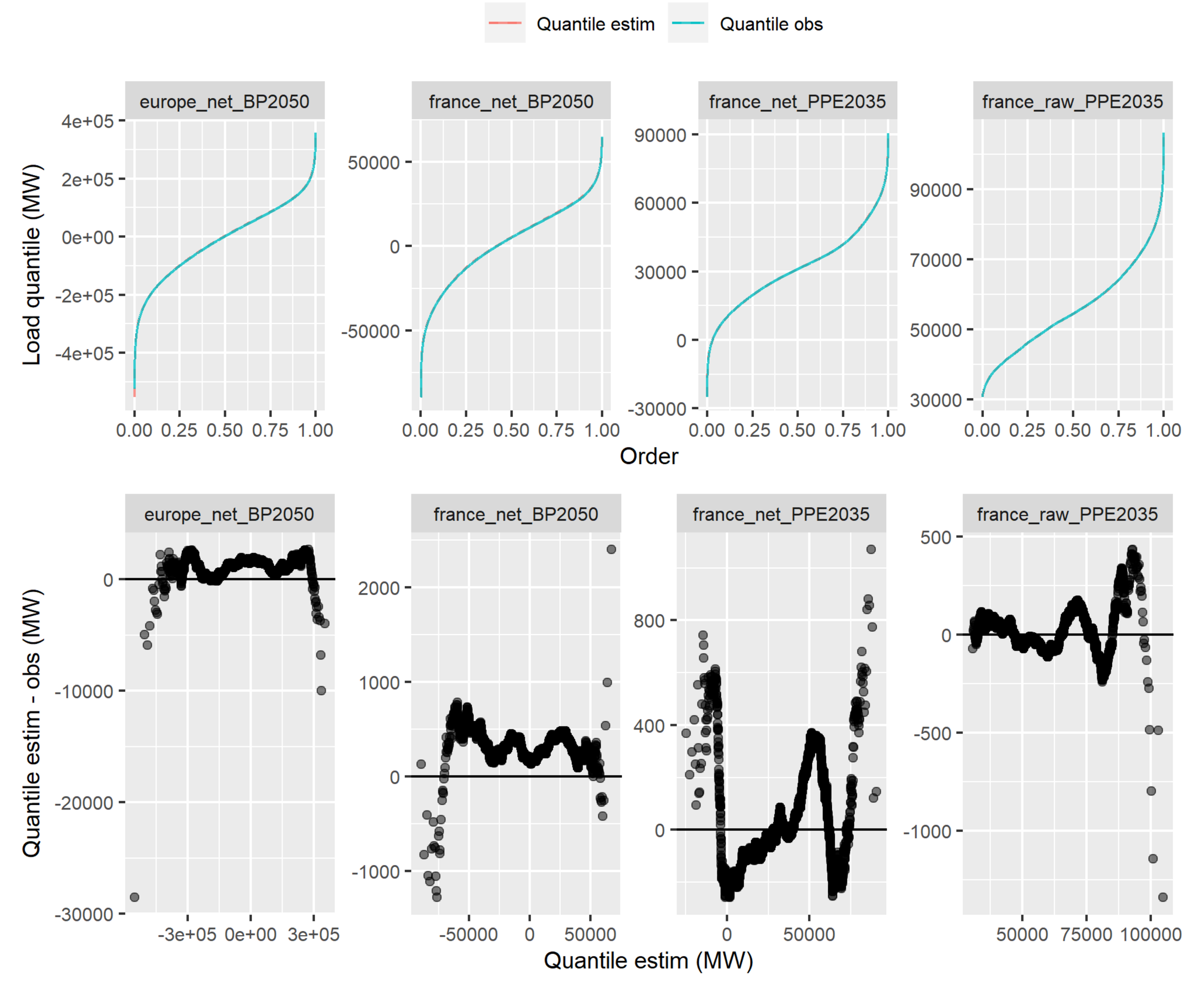

Figure 4: Observed (200 years) and estimated quantiles from a satisfying sample.

Figure 5: Scores of a satisfying sample vs. average scores of random samples.

Discussion

- Definitions of distance between years ?
- Avoiding bad sampling vs. minimization ?
- Number of scores ?
- New variables with spatial or temporal aggregation ?
- Definitions of score threshold M_i ?
- Sufficient number of climate models ? Sufficient number of socio-economic scenarios ?
- Sampling strategies ?

Conclusion

- Lower approximation error with satisfying samples of years.
- Experiments on weather and energy data.
- A few R packages: fields, energy, lubridate, posterdown.
- Code : https://github.com/rte-france/scenclimsample